Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages

Siyu Liang and **Gina-Anne Levow**
University of Washington
liangsy, levow@uw.edu

Abstract

The development of Automatic Speech Recognition (ASR) has yielded impressive results, but its use in linguistic fieldwork remains limited. Recordings collected in fieldwork contexts present unique challenges, including spontaneous speech, environmental noise, and severely constrained datasets from underdocumented languages. In this paper, we benchmark the performance of two fine-tuned multilingual ASR models, MMS and XLS-R, on five typologically diverse low-resource languages with control of training data duration. Our findings show that MMS is best suited when extremely small amounts of training data are available, whereas XLS-R shows parity performance once training data exceed one hour. We provide linguistically grounded analysis for further provide insights towards practical guidelines for field linguists, highlighting reproducible ASR adaptation approaches to mitigate the transcription bottleneck in language documentation.

1 Introduction

Automatic Speech Recognition (ASR) has achieved significant breakthroughs in recent years, with deep learning-based models reported to reach near-human word error rates for high-resource languages (Radford et al., 2023; Baevski et al., 2020). However, these advancements have largely been driven by massive transcribed datasets (e.g. Chang et al. (2022); Panayotov et al. (2015); Godfrey et al. (1992)), leaving a substantial performance gap for low-resource languages, particularly those encountered in linguistic fieldwork (Guillaume et al., 2022a). Fieldwork speech data presents distinct challenges, including spontaneous speech, varied recording setups, and typologically diverse linguistic features, all of which could degrade the performance of ASR models trained on standardized speech corpora.

Linguistic fieldwork plays a critical role in preserving endangered languages and documenting linguistic diversity. These recordings capture not only the linguistic structures of a language, but also oral traditions, discourse patterns, and sociolinguistic variations (Himmelmann, 1998; Austin and Sallabank, 2011). However, while some well-researched low-resource languages have substantial datasets (Guillaume et al., 2022b; Przezdziak, 2024), there is usually limited data to bootstrap an ASR model for most field linguists. Evaluations of the ASR approaches usually tend to focus on one language (Jones et al., 2024; Rijal et al., 2024; Guillaume et al., 2022b; Mainzinger and Levow, 2024) or are inconsistent regarding data size, genre, etc. in the sample (Jimerson et al., 2023). Evaluations of models for low-resource languages also tend to favor clean, good quality, read speech (Rijal et al., 2024; Mainzinger and Levow, 2024; Jimerson et al., 2023), compared with the noisier and more spontaneous speech of fieldwork recordings.

1.1 The transcription bottleneck

Linguistic fieldwork plays a crucial role in documenting endangered and under-researched languages, yet the process of manually transcribing recordings remains a significant barrier: transcribing a single hour of audio in a newly documented language can require up to 50 hours of work (Shi et al., 2021). Moreover, many of the fieldwork languages also lack standardized orthographies, requiring a handful of trained linguists to make discerning decisions during transcription. As a result, the volume of untranscribed linguistic data continues to grow, creating a severe bottleneck in language documentation, analysis, and distribution (Anastasopoulos and Chiang, 2018; Bird, 2020; Thieberger, 2012).

The dependence on large transcribed datasets for training ASR models exacerbates this issue, as most endangered and low-resource languages

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 26–37

lack sufficient annotated speech data to support model development (Levow et al., 2021). Without adequate transcriptions, traditional supervised ASR methods remain ineffective, requiring alternative approaches that can leverage limited data more efficiently (Dunbar et al., 2019; Baevski et al., 2020, 2021).

1.2 ASR for Low-resource Data

The application of ASR in linguistic fieldwork closely parallels the development of low-resource ASR research. Since the 2010s, much of this work has focused on languages from the IARPA Babel project, which served as a cornerstone for ASR development in low-resource settings (Miao et al., 2013; Cui et al., 2014; Grézl et al., 2014). Research leveraging Babel datasets introduced key techniques such as transfer learning, multilingual adaptation, and data augmentation, which have since become fundamental to ASR advancements in underdocumented languages (Zhang et al., 2014; Khare et al., 2021; Vanderreydt et al., 2022; Guillaume et al., 2022a).

The widespread adoption of the Kaldi toolkit (Povey et al., 2011) further propelled ASR research in these domains, enabling the development of reproducible pipelines and fostering the open distribution of Kaldi-compatible datasets (Yadava and Jayanna, 2017; Milde and Köhn, 2018; Adams et al., 2021; Zhang et al., 2022). Concurrently, researchers have explored approaches such as transfer learning and fine-tuning from multilingual pre-trained models (Guillaume et al., 2022a; Sikasote and Anastasopoulos, 2021) or adapting English-centric models to new linguistic domains (Kim et al., 2021; Thai et al., 2020). Additionally, self-supervised and semi-supervised learning approaches have gained traction as viable solutions for overcoming transcription scarcity, further bridging the gap between ASR and field linguistics (Babu et al., 2021; Baevski et al., 2021).

1.3 Fine-tuning Pre-trained ASR Models

Fine-tuning pre-trained ASR models has emerged as a key approach for improving recognition accuracy in low-resource settings, particularly for linguistic fieldwork recordings (Guillaume et al., 2022a; Pillai et al., 2024; Nowakowski et al., 2023). Self-supervised learning models, such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), MMS (Massive Multilingual Speech) Model (Pratap et al., 2024), and XLS-R (Babu et al., 2021), have demonstrated the ability to learn generalized speech representations from large-scale multilingual datasets, significantly reducing the need for extensive transcriptions in under-documented languages. Studies on specific low-resource languages, such as Bribri (Coto-Solano, 2021), Japhug (Guillaume et al., 2022b), Mvskoke (Mainzinger and Levow, 2024), and the Čakavian dialect of Croatian (Jones et al., 2024), underscore the benefits of adapting large multilingual models and report considerable reductions in error rates even with very limited data.

Nevertheless, recent work suggests that no single architecture or end-to-end approach consistently outperforms others under extremely low-resource conditions (Jimerson et al., 2023). Some studies advocate experimenting with multiple toolkits and hyperparameter configurations to identify solutions best suited to the language at hand. Indeed, while fully fine-tuning massive models can be effective, it often requires large amounts of computational resources, can risk overfitting with very small datasets, and demands updating millions of parameters.

Instead of modifying all model parameters, adapters introduce small trainable layers while keeping the base pre-trained model frozen, thereby reducing memory requirements and improving efficiency (Houlsby et al., 2019). This makes them particularly useful for linguistic fieldwork applications, where data is scarce and computational resources are limited. The MMS model developed by Meta (Pratap et al., 2024) integrates adapter layers specifically designed for ASR, enabling efficient adaptation to new languages with minimal training data. Studies in low-resource settings (Bai et al., 2024; Mainzinger and Levow, 2024) have shown that adapter-based fine-tuning can achieve performance comparable to full fine-tuning while requiring significantly fewer trainable parameters. By avoiding overfitting on small datasets and focusing on the most relevant parameters for language adaptation, adapter-based methods offer an attractive balance between accuracy and efficiency, an approach increasingly vital to sustaining language documentation efforts in the face of extremely sparse resources.

In contrast to earlier studies that often focus on a single language or on clean, scripted corpora, our work systematically evaluates both MMS and XLS-R in truly low-resource fieldwork conditions spanning multiple typologically diverse languages.

By examining noise-heavy, spontaneous recordings rather than controlled speech, we test model adaptability in settings that more accurately reflect real-world linguistic documentation. Further, our fine-grained error analysis explores how each model handles the nuanced phonological features that typify endangered and under-documented languages—details that have often been overlooked in prior research. Together, these innovations provide a clearer roadmap for linguists seeking practical ASR solutions under extreme data scarcity and diverse orthographic conventions.

2 Data

We test the performance of fine-tuned ASR models on five typologically varied low-resource languages: Cicipu (ISO639-3: awc, McGill (2012)), Mocho' (ISO639-3: mhc, Pérez González (2018)), Toratán (ISO639-3: rth, (Jukes, 2010)), Ulwa (ISO639-3: yla, (Barlow, 2018a)), and Upper Napo Kichwa (ISO639-3: quw, (Grzech, 2020)). These languages span multiple language families and exhibit distinct phonetic, phonological, and morphological features. The data is drawn from the Endangered Languages Archive (ELAR)[1], where gold-standard transcriptions can be derived from the recordings and the corresponding time aligned transcriptions in the ELAN (Brugman and Russel, 2004) format. The dataset encompasses a variety of genres, such as greetings, narratives, ritual discourse, interviews, elicitation sessions, folktales, and cultural practices, the details of which are given in Table 5 of Appendix B. Table 1 provides an overview of key linguistic features, including vowel and consonant inventories as well as tonal systems.

2.1 Data details

Given that the recordings were made in naturalistic fieldwork environments, they exhibit acoustic idiosyncrasies that could pose significant challenges for ASR. There's background noise from outdoor settings, such as wind, animals, and community sounds, in many of the recordings. We also observe code-switching, usually in the regional dominant languages. For example, Toratán speakers are also speakers of Manado Malay (with various degrees of fluency), and loans from Malay are generally not adapted to Toratán phonology (Himmelmann and Wolff, 1999). We also observe significant Spanish code-mixing in Mocho', as well as some English content in Cicipu.

The dataset sizes vary, with archived speech ranging from approximately 2 to 22 hours per language. However, not all archived data have been transcribed. This variation reflects real-world constraints in linguistic fieldwork, where some languages have more extensive documentation than others.

2.2 Dataset Pre-processing

All recordings were resampled to 16 kHz (the training sampling rate for both MMS and XLS-R), converted to mono-channel WAV format, and aligned with their corresponding transcriptions. During transcription pre-processing, we referenced the phonological description of each language to ensure that punctuation marks or special characters used to denote phonological features were retained (see Table 5 of Appendix B). Audio segments explicitly transcribed as non-linguistic sounds, such as laughter, were excluded from the dataset. We also removed utterances that contain only filler words, such as 'mhm', 'aaa', etc. A breakdown of the audio lengths before and after pre-processing is given in Table 6 of Appendix B

We created four total train+dev duration configurations for each language—10, 30, 60, and 120 minutes—before splitting the data into training (90%) and development (10%) sets. In addition, we set aside a fixed 10-minute test set for final evaluation. To maintain consistency and facilitate interpretation, larger dataset splits were structured as supersets of smaller ones. We did not designate a held-out speaker, as field linguists typically work with a limited number of consultants and would prioritize consistent model performance across familiar speakers (Liu et al., 2023). Details of the cleaned dataset are shown in Table 2. The last column of the table lists the number of unique characters used in the language. Due to different transcription conventions, features such as nasalization, vowel length and voicing could be indicated with diacritics or extra letters. Therefore, although transcriptions are meant to be phonemic, the number of unique characters might not match the number of contrastive vowels and consonants. In the case of Cicipu, its unusually large inventory of 93 characters is due to the number of all possible combinations of nasality, tone, and vowel quality marking.

[1] https://www.elararchive.org/

Language	Family	Region	#V	#C	#T	Features
Cicipu	Niger-Congo	Nigeria	28	27	4	Ṽ, Vː, Cː
Mocho'	Mayan	Mexico	10	27	2	Vː
Toratán	Austronesian	Indonesia	5	21	0	
Ulwa	Keram	Papua New Guinea	8	13	0	[-voice, +son]
Upper Napo Kichwa	Quechuan	Ecuador	8	20	0	Vː

Table 1: Linguistic data used for the study, showing language family, region spoken, phoneme inventory size, tones, and phonological features such as nasality, vowel length, consonant gemination, etc.

Language	#Spk	Avg. Leng.	#Char
Cicipu	33	2.1	93
Mocho'	6	2.0	29
Toratán	13	2.35	27
Ulwa	6	3.65	25
U.N. Kichwa	16	3.79	33

Table 2: Dataset statistics for different languages, including the number of speakers, average utterance length in seconds, and character inventory.

3 Methodology

This study investigates the effectiveness of fine-tuning multilingual ASR models to address the unique challenges posed by low-resource linguistic fieldwork recordings. By evaluating the performance of two state-of-the-art models, we aim to determine how fine-tuning can enhance recognition accuracy on typologically diverse, low-resource languages. In addition, we discuss the impact of key factors, including training data size, model choice, and pre-trained model features, to provide practical insights for ASR adaptation in fieldwork contexts.

3.1 Models

Our goal is to fine-tune state-of-the-art multilingual ASR models that have been pre-trained on large-scale speech corpora. Specifically, we evaluate models from Meta's Massively Multilingual Speech (MMS) project (Pratap et al., 2024) alongside XLS-R (Babu et al., 2021), a widely used multilingual ASR model.

MMS is based on the wav2vec 2.0 framework (Baevski et al., 2020), which employs self-supervised learning to extract generalized speech representations from vast amounts of unlabeled audio. The model has been trained on 1,406 languages, making it one of the most comprehensive ASR models for multilingual speech recognition. Specifically, we choose MMS-1B-l1107, a 1-billion parameter model fine-tuned specifically for ASR with an additional 2-million parameter adapter, which supports ASR in 1107 languages out of the box (Houlsby et al., 2019). The adapter facilitates efficient language-specific fine-tuning while preserving the generalized multilingual knowledge encoded in the base model.

XLS-R (Babu et al., 2021) is a multilingual ASR model pre-trained on 128 languages using the wav2vec 2.0 framework. It has been extensively used for low-resource ASR and cross-lingual transfer learning, making it a strong baseline for evaluating ASR performance in linguistic fieldwork settings. Unlike MMS, which is trained on over 1,000 languages, XLS-R has been optimized for a balanced selection of 128 languages, with a strong focus on phonetic diversity. This makes it particularly useful for comparison against MMS models to assess the effectiveness of scaling multilingual pre-training to extremely low-resource languages. The XLS-R-300m with 300 million parameters is chosen for the study.

None of the languages used in the study, with the exception of Upper Napo Kichwa, is represented in the training data of the two models. A discussion of the possible effects is included in Section 4.2.

3.2 Implementation

Following the fine-tuning procedure outlined by von Platen[2] and using the Hugging Face Transformers library (Wolf et al., 2020), we implement model-specific strategies. For MMS-1B-l1107, only the adapter layers are fine-tuned, with the base model frozen. For XLS-R, the entire model is fine-tuned. To assess the effect of data size, models are trained on subsets of 10, 30, 60, and 120 minutes of transcribed fieldwork data per language. Early stopping, based on the development set Character Error Rate (CER), mitigates overfitting. Hyperparameter details, including batch size, learning rate, opti-

[2] https://huggingface.co/blog/mms_adapters

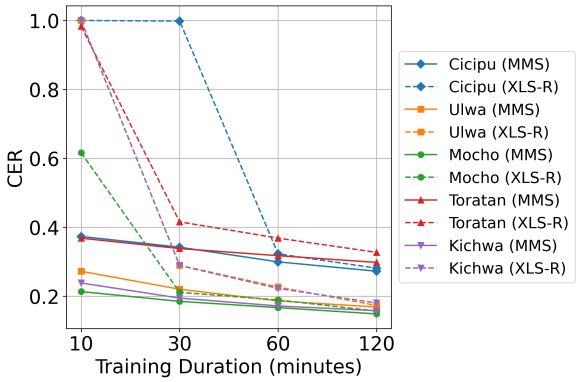

Figure 1: CER comparison for MMS-1b1107 and XLS-R-300m models across five languages. The MMS model performs markedly better under extremely low-resource settings (less than 1 hour), but XLS-R performs similarly well with 2 hours of data. Points are connected to aid trend reading and do not imply performance at intermediate durations.

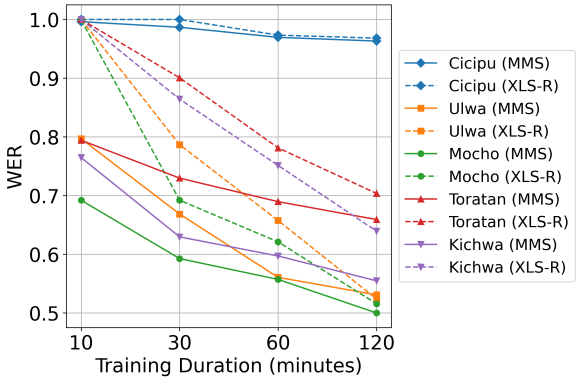

Figure 2: WER comparison for MMS-1b1107 and XLS-R-300m models across five languages.

mization strategy, and training details are provided in Appendix A.

4 Results

Overall, the performance of the MMS and XLS-R models in automatic speech recognition (ASR) tasks on low-resource language data is comparable, though nuanced differences emerge depending on the availability of training data. In general, the MMS model outperforms XLS-R under extremely low-resource conditions (i.e., less than one hour of transcribed data). However, XLS-R demonstrates a marked improvement as the size of the training data scales beyond this threshold, ultimately becoming on par with MMS. Figure 1 and Figure 2 illustrate these trends in Character Error Rate (CER) and Word Error Rate (WER) metrics across the five languages in this study.

In our analysis, CER serves as the primary evaluation metric. Unlike WER, which operates at the word level and often presupposes well-defined word boundaries and stable orthographic forms, CER better reflects the needs of field linguists, who often lack enough data to be able to use language models. In fieldwork contexts, the primary goal of transcription is to capture phonetic accuracy, particularly in languages without standardized orthographies. During model training, the best model metric is set to CER to ensure that model performance aligns with the core task of producing reliable phoneme-level transcriptions from spontaneous and noisy speech data.

4.1 Data size effect

The relationship between training data size and model performance is critical in understanding model suitability for field linguistics applications. For both MMS and XLS-R, performance improvements start to plateau when training data exceeds approximately one hour. This finding suggests steady although diminishing returns beyond this point, aligning with previous observations in low-resource ASR research (Guillaume et al., 2022a).

In scenarios where less than one hour of data is available, MMS consistently achieves lower error rates, likely due to its extensive multilingual pre-training on over 1,000 languages. For field linguists dealing with extremely limited resources, we thus recommend fine-tuning MMS to achieve acceptable ASR performance. However, when approximately one hour of data or more is obtainable, XLS-R becomes a more effective option due to its improving performance with increasing data volumes. This suggests that one hour of transcribed data serves as a practical threshold for developing a robust fine-tuned ASR system in fieldwork contexts.

It is worth noting that Cicipu exhibits particularly high error rates under extreme data scarcity, including a character error rate approaching 1.0 at 10 minutes of data for both models and at 30 minutes for XLS-R. Cicipu's unusually large orthographic inventory (93 unique characters reflecting combinations of nasality, tone, and vowel quality) requires more training examples to accurately learn the mapping from acoustics to graphemes. Consequently, with only a few minutes of labeled data, neither model can fully learn Cicipu's complex phonological and orthographical features.

4.2 MMS vs. XLS-R

The MMS model excels in settings with minimal data due to its multilingual pretraining on a vast corpus that includes low-resource languages. Moreover, unlike XLS-R—whose core version is primarily self-supervised and not initially fine-tuned for ASR—MMS has already undergone a large-scale ASR fine-tuning step. This means that MMS starts off with more task-specific parameters, making it more effective than XLS-R in extremely low-data regimes. However, MMS's reliance on primarily read speech data (e.g. Bible translations) may limit its adaptability to spontaneous speech environments, which are common in linguistic fieldwork recordings.

In contrast, XLS-R benefits from a more diverse training corpus that encompasses conversational and spontaneous speech, allowing it to generalize better once sufficient data becomes available. Indeed, Mainzinger and Levow (2024) reported superior performance of MMS over XLS-R when fine-tuning Mvskoke—likely due to both the advantage of MMS's ASR fine-tuning and the fact that much of the Mvskoke training material was similarly read or scripted speech.

Since several related dialects of Kichwa (Eberhard et al., 2024, 2025) were included in the MMS pre-training dataset, we investigated whether the performance gap between MMS and XLS-R would be *larger* for Kichwa than for the other languages in our study. Specifically, we fit a linear mixed-effects model (with random intercepts for each language) to our character error rate (CER) data, using *model* (MMS vs. XLS-R), *time*, and an indicator *similar* (1 = Kichwa, 0 = other languages) as fixed effects. If Kichwa had benefitted disproportionately from MMS's pre-training, we would have observed a significant positive interaction in the model. However, the interaction term (*model* $\times$ *similar*) was small and not statistically significant ($\beta = 0.021$, $p = 0.896$), indicating that while MMS outperforms XLS-R overall, the additional advantage for Upper Napo Kichwa is not discernibly greater than for the other languages.

Further research is needed to evaluate whether the performance trend continues with larger datasets, particularly for languages with similar phonological and morphological complexity as those in this study. Additionally, the effectiveness of adapter-based fine-tuning for MMS suggests that optimizing model architecture for scalable adaptation could yield further improvements.

Lang	Model	Tone	Nas.	V-Len	C-Len
Cicipu	MMS	0.199	0.337	0.239	0.174
	XLS-R	0.215	0.337	0.248	0.183
Mocho'	MMS	—	—	0.154	—
	XLS-R	—	—	0.160	—

Table 3: Phonological error rates (0–1 scale) for Cicipu and Mocho'. Cicipu shows higher confusion in tone, nasality, and consonant length, while Mocho' displays more issues with vowel length. Both models (MMS vs. XLS-R) yield broadly similar error patterns.

4.3 Error analysis

We performed a phonologically informed error analysis on two of the languages in our dataset, Cicipu and Mocho', both of which exhibit segmental contrasts that could be challenging for ASR models. Cicipu's orthography explicitly marks tone and nasality with diacritics and differentiates both vowel and consonant length with doubled letters (e.g., 'aa' for long vowels, 'tt' for geminate consonants), making it well suited for evaluating the system's performance on these phonological categories. Mocho' similarly features a vowel length distinction encoded with doubled vowel letters. Other languages in our dataset, such as Ulwa and Upper Napo Kichwa, contain too few instances of long vowels or voiceless sonorants for a robust category-based analysis.

4.3.1 Error rates

To quantify performance on these features, we leverage character-level alignments (details in Appendix C) and calculate phonological segment error rates (Table 3). For each category $C \in \{\text{Tone}, \text{Nasality}, \text{V\_length}, \text{C\_length}\}$, we sum all substitutions, deletions, and insertions that affect that category and normalize by the total number of reference tokens L_C exhibiting C.

As shown in Table 3, both MMS and XLS-R struggle with Cicipu's tone, nasality, and consonant length, each exhibiting error rates in the range of 30–38%. By contrast, vowel-length confusion is comparatively low (7–9%). For Mocho', the long–short vowel distinction remains problematic, with error rates around 35–38%. These findings suggest that neither model has a strong advantage for these particular phonological categories; even after fine-tuning, nuanced contrasts such as nasality and tone remain challenging.

Lang	Category	S%	D%	I%
Cicipu	Tone	47.62	32.11	20.27
	Nasality	0.00	70.00	30.00
	V-Len	55.58	27.86	16.56
	C-Len	44.18	33.90	21.92
Mocho'	V-Len	58.29	26.86	14.86

Table 4: Percentage of substitutions (S%), deletions (D%), and insertions (I%) in XLS-R 120 min output, for each phonological category in Cicipu and Mocho'.

4.3.2 Error distribution

For further investigation, we performed a more detailed error analysis on the output from the 120-minute model of XLS-R for Cicipu and Mocho'. Table 4 breaks down each category by the percentage of substitutions (*S*), deletions (*D*), and insertions (*I*).

In Cicipu, nearly half of the tone errors ($\sim 48\%$) are substitutions, indicating confusion over which tone mark to apply, while around a third are deletions and the remainder insertions. Nasality errors, by contrast, skew heavily toward deletions ($\sim 70\%$), suggesting the model often fails to detect nasal vowel features. Substitutions are rare for nasality, indicating that the system either omits it entirely or adds it spuriously rather than confusing it with another tone diacritic. Vowel-length errors ($\sim 44\%$ for deletions and insertions) and consonant-length errors ($\sim 55\%$ for deletions and insertions) reflect a high level of segment-level confusion, whereas confusion with a different vowel ($\sim 56\%$ substitutions) or consonant ($\sim 44\%$ substitutions) is also frequent. For Mocho', vowel-length confusion is likewise dominated by substitutions ($\sim 58\%$), a pattern similar to Cicipu which reveals that XLS-R often misidentifies one segment in the long vowel. The distributions point out that while the model does capture some acoustic correlates of nasality, tone, and length, it nevertheless struggles to map them consistently to diacritics and extended graphemes in low-resource scenarios.

Overall, these patterns highlight the persistent challenge of representing languages with complex orthographies and rich phonological inventories. Even after multilingual pre-training and fine-tuning, contrasts such as tone or nasality may be overlooked when the amount of transcribed data is minimal. Addressing these gaps may require linguistically informed data augmentation, specialized adapter modules, or loss functions that explicitly emphasize distinct phonological categories. In extremely low-resource settings, such targeted methods could provide the additional examples and acoustic cues needed for more accurate transcription of endangered languages.

5 Conclusion

Our experiments show that fine-tuned multilingual ASR models can substantially reduce the transcription burden for endangered and low-resource languages. Across five typologically diverse languages, MMS proved more effective with extremely limited labeled data, whereas XLS-R caught up once approximately one hour of transcribed material was available. By using Character Error Rate (CER) rather than Word Error Rate (WER), we focus on phoneme-level accuracy—a more direct measure for languages without standardized orthographies. Despite improvements in overall accuracy, both models struggled with challenging phonological categories in Cicipu, such as tone and consonant length, and exhibited a high rate of vowel-length confusion in Mocho'. These findings confirm that current multilingual ASR systems are indeed helpful for language documentation but still require targeted adaptations to handle nuanced phonological contrasts in under-resourced settings.

6 Future Work

Although our study confirms that fine-tuning multilingual ASR models can substantially reduce transcription overhead for low-resource languages, several research directions remain promising for further performance gains. One compelling approach is continued pre-training (CoPT) on unlabeled in-language audio. DeHaven and Billa (2022) show that CoPT on a wav2vec 2.0–based multilingual model can match or outperform pseudo-labeling techniques while being more computationally efficient. Similarly, Nowakowski et al. (2023) demonstrate that CoPT on about 234 hours of Sakhalin Ainu audio yields a considerable reduction in error, beyond what standard multilingual fine-tuning achieves. CoPT has also proven effective in domain adaptation, especially for noisy data or new speaker types (Attia et al., 2024). While the scarcity of fieldwork data could limit the scale of CoPT, even incremental benefits may substantially ease the manual transcription effort.

A second avenue is leveraging diverse augmentation methods to enlarge the effective training set. Self-training (pseudo-labeling) uses an initial

ASR model to generate transcripts for unlabeled audio, which can then be added to the training pool. This method has consistently boosted low-resource ASR performance (Bartelds et al., 2023), particularly when coupled with filtering or iterative refinement. TTS-based augmentation offers another option: if a target-language text-to-speech system is available, synthesizing speech from text yields additional "perfectly labeled" data, potentially improving recognition robustness (ibid). Finally, common audio perturbations, speed/pitch changes, SpecAugment, and noise injection, remain valuable for avoiding overfitting and preparing the model for real-world variability.

A final challenge involves better capturing difficult phonological categories, such as tone, nasality, and consonant length. Adapters in MMS could be extended or reconfigured to emphasize language-specific features, while training regimes could incorporate acoustic or phonological priors explicitly. Future work might integrate fine-grained linguistic annotations (if available) or employ specialized masking strategies during CoPT to boost the model's sensitivity to subtle contrasts. Combining these techniques into user-friendly toolkits will be essential for widespread adoption by field linguists, who often have limited computational resources yet require high-accuracy, phoneme-level transcriptions for documenting and revitalizing endangered languages.

7 Limitations

Despite promising results, several specific limitations affect the generalizability and applicability of our approach. The most critical limitation is related to the data size and representativeness of the linguistic diversity considered. Our study focused on a small number of typologically diverse languages, each with relatively limited datasets ranging from just a few minutes to two hours. As such, the models' performances may not generalize to other endangered or low-resource languages with distinct phonological or orthographic features.

Additionally, due to constraints inherent in linguistic fieldwork, the training and test datasets often contained data from the same speakers, potentially inflating model accuracy estimates. Future research should validate these findings with genuinely held-out speakers to better gauge model robustness to speaker variability.

Moreover, the orthographic inconsistencies and the absence of standardized orthographies in our datasets likely influenced model performance, especially for phonologically complex categories like tone, vowel length, and nasality. This issue highlights a broader limitation: ASR models trained under these conditions may struggle to generalize to spontaneous and noisy field recordings, especially when orthographic conventions vary within and across datasets.

Finally, computational resource limitations (training on a single NVIDIA T4 GPU with constrained runtimes) restrict our ability to fine-tune larger models or extensively optimize hyperparameters, which may have further improved performance. Addressing these limitations would require additional computational resources and potentially more extensive data augmentation strategies tailored explicitly to low-resource linguistic contexts.

Acknowledgments

We are grateful to the depositors and leaders of the Endangered Languages Archive (ELAR, http://elararchive.org) for sharing their invaluable resources which made this project possible.

References

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. 2021. User-friendly Automatic Transcription of Low-resource Languages: Plugging ESPnet into Elpis. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 51–62, Online. Association for Computational Linguistics.

Antonis Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. *arXiv preprint*. ArXiv:1803.08991 [cs].

Ahmed Adel Attia, Dorottya Demszky, Tolulope Ogunremi, Jing Liu, and Carol Espy-Wilson. 2024. Continued Pretraining for Domain Adaptation of Wav2vec2.0 in Automatic Speech Recognition for Elementary Math Classroom Settings. *arXiv preprint*. ArXiv:2405.13018 [cs] version: 1.

Peter K. Austin and Julia Sallabank. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge University Press. Google-Books-ID: 0XZRauYgO6AC.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika

Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs].

Alexei Baevski, Wei-Ning Hsu, Alexis CONNEAU, and Michael Auli. 2021. Unsupervised Speech Recognition. In *Advances in Neural Information Processing Systems*, volume 34, pages 27826–27839. Curran Associates, Inc.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Junwen Bai, Bo Li, Qiujia Li, Tara N. Sainath, and Trevor Strohman. 2024. Efficient Adapter Finetuning for Tail Languages in Streaming Multilingual ASR. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10841–10845. ISSN: 2379-190X.

Russell Barlow. 2018a. Documentation of Ulwa, an endangered language of Papua New Guinea.

Russell Barlow. 2018b. A Grammar of Ulwa.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. *arXiv preprint*. ArXiv:2305.10951 [cs].

Steven Bird. 2020. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Hennie Brugman and Albert Russel. 2004. Annotating Multi-media / Multi-modal resources with ELAN.

Xuankai Chang, Takashi Maekaku, Yuya Fujita, and Shinji Watanabe. 2022. End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation. *arXiv preprint*. ArXiv:2204.00540 [cs].

Rolando Coto-Solano. 2021. Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Xiaodong Cui, Brian Kingsbury, Jia Cui, Bhuvana Ramabhadran, Andrew Rosenberg, Mohammad Sadegh Rasooli, Owen Rambow, Nizar Habash, and Vaibhava Goel. 2014. Improving deep neural network acoustic modeling for audio corpus indexing under the IARPA babel program. In *Interspeech 2014*, pages 2103–2107. ISCA.

Mitchell DeHaven and Jayadev Billa. 2022. Improving Low-Resource Speech Recognition with Pretrained Speech Models: Continued Pretraining vs. Semi-Supervised Training. *arXiv preprint*. ArXiv:2207.00659 [cs].

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2019. The Zero Resource Speech Challenge 2019: TTS without T. *arXiv preprint*. ArXiv:1904.11469 [cs].

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Napo Quichua. Edition: 27 Publisher: SIL International.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. Tena Lowland Quichua. Edition: 26 Publisher: SIL International.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1. ISSN: 1520-6149.

Karolina Grzech. 2020. Upper Napo Kichwa: a documentation of linguistic and cultural practices.

Frantisek Grézl, Martin Karafiát, and Karel Veselý. 2014. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658. ISSN: 2379-190X.

Séverine Guillaume, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyen, Maxime Fily, Guillaume Jacques, and Alexis Michaud. 2022a. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. pages 4905–4909. International Speech Communication Association.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyen, and Maxime Fily. 2022b. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. 36(1):161–196. Publisher: De Gruyter Mouton Section: Linguistics.

Nikolaus P Himmelmann and John U Wolff. 1999. *Toratán (Ratahan)*, volume 130. Lincom Europa.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR. ISSN: 2640-3498.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Austin Jones, Shulin Zhang, John Hale, Margaret Renwick, Zvjezdana Vrzic, and Keith Langston. 2024. Comparing Kaldi-Based Pipeline Elpis and Whisper for Čakavian Transcription. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 61–68, Bangkok, Thailand. Association for Computational Linguistics.

Anthony Jukes. 2010. Documentation of Toratán (Ratahan).

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Interspeech 2021*, pages 1529–1533. ISCA.

Jiyeon Kim, Mehul Kumar, Dhananjaya Gowda, Abhinav Garg, and Chanwoo Kim. 2021. Semi-supervised transfer learning for language expansion of end-to-end speech recognition models to low-resource languages. *arXiv preprint*. ArXiv:2111.10047 [eess].

Gina-Anne Levow, Emily P. Ahn, and Emily M. Bender. 2021. Developing a Shared Task for Speech Processing on Endangered Languages. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1:96–106.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-Tuning ASR models for Very Low-Resource Languages: A Study on Mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Stuart McGill. 2012. Cicipu documentation.

Stuart McGill. 2014. Cicipu. *Journal of the International Phonetic Association*, 44(3):303–318.

Yajie Miao, Florian Metze, and Shourabh Rawat. 2013. Deep maxout networks for low-resource speech recognition. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 398–403.

Benjamin Milde and Arne Köhn. 2018. Open Source Automatic Speech Recognition for German. *arXiv preprint*. ArXiv:1807.10311 [cs].

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting Multilingual Speech Representation Model for a New, Underresourced Language through Multilingual Fine-tuning and Continued Pretraining. *Information Processing & Management*, 60(2):103148. ArXiv:2301.07295 [cs].

Erin O'Rourke and Tod D. Swanson. 2013. Tena Quichua. *Journal of the International Phonetic Association*, 43(1):107–120. Publisher: Cambridge University Press.

Naomi Elizabeth Palosaari. 2011. *Topics in Mocho' phonology and morphology*. Ph.D., The University of Utah, United States – Utah. ISBN: 9781124576213.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. ISSN: 2379-190X.

Leena G. Pillai, Kavya Manohar, Basil K. Raju, and Elizabeth Sherly. 2024. Multistage Fine-tuning Strategies for Automatic Speech Recognition in Low-resource Languages. *arXiv preprint*. ArXiv:2411.04573 [cs].

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Agnieszka Przezdziak. 2024. *Optimizing Speech Recognition for Low-Resource Languages: Northern Sotho*.

Jaime Pérez González. 2018. Documentation of Mocho' (Mayan): Language Preservation through Community Awareness and Engagement.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR. ISSN: 2640-3498.

Sanjay Rijal, Shital Adhikari, Manish Dahal, Manish Awale, and Vaghawan Ojha. 2024. Whisper Finetuning on Nepali Language. *arXiv preprint*. ArXiv:2411.12587 [cs].

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec. *arXiv preprint*. ArXiv:2101.10877 [eess].

Claytone Sikasote and Antonios Anastasopoulos. 2021. BembaSpeech: A Speech Recognition Corpus for the Bemba Language. *arXiv preprint*. ArXiv:2102.04889 [cs].

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully Convolutional ASR for Less-Resourced Endangered Languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 126–130, Marseille, France. European Language Resources association.

Nick Thieberger. 2012. *The Oxford Handbook of Linguistic Fieldwork*. OUP Oxford. Google-Books-ID: 86AE2_0nPbkC.

Geoffroy Vanderreydt, François Remy, and Kris Demuynck. 2022. Transfer Learning from Multi-Lingual Speech Translation Benefits Low-Resource Speech Recognition. In *Interspeech 2022*, pages 3053–3057. ISCA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Michael Wroblewski. 2012. Amazonian Kichwa Proper: Ethnolinguistic Domain in Pan-Indian Ecuador. *Journal of Linguistic Anthropology*, 22(1):64–86. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1548-1395.2012.01134.x.

G. Thimmaraja Yadava and H S Jayanna. 2017. Development and comparison of ASR models using kaldi for noisy and enhanced kannada speech data. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1832–1838. Conference Name: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) ISBN: 9781509063673 Place: Udupi Publisher: IEEE.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition. *arXiv preprint*. ArXiv:2110.03370 [cs].

Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. 2014. Improving deep neural network acoustic models using generalized maxout networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 215–219. ISSN: 2379-190X.

A Training Details

The models were trained on an NVIDIA T4 GPU, with training times ranging from approximately 1 to 60 minutes per model. The hyper-parameters were defined as follows:

- **Learning rate:** MMS: 1e-3; XLS-R: 3e-4
- **Maximum epochs:** 30
- **Best model metric:** Character Error Rate (CER)
- **Early stopping:** 3 epochs
- **Early stopping threshold:** 0.003

B Dataset Details

Table 5 provides detailed information on the genres and types of content present in the datasets used in this study, along with key linguistic references and citations to the original documentation archives. Table 6 summarizes the total archived hours of audio recordings for each language and the amount of data remaining after the cleaning and preprocessing steps described earlier in the manuscript.

Language	Genres of content	Phonological Description	Documentation
Cicipu	Greetings, conversations, hortative discourse, narratives, procedural discourse, ritual discourse, elicitation activities	McGill (2014)	McGill (2012)
Mocho'	Biographical and non-biographical narratives (historical events, myths, local beliefs, traditional building, witchcraft), prayer, conversation, elicitation sessions, text translation	Palosaari (2011)	Pérez González (2018)
Toratán	Conversational data, elicitation sessions, narratives (personal history, folk tales)	Himmelmann and Wolff (1999)	Jukes (2010)
Ulwa	Conversational data, traditional stories, personal stories, traditional singing and dancing video	Barlow (2018b)	Barlow (2018a)
Upper Napo Kichwa	Grammatical elicitation, life interviews	Wroblewski (2012), O'Rourke and Swanson (2013)	Grzech (2020)

Table 5: Details of depository content for languages used in this paper, related linguistic work referenced, and original documentation citations.

Language	Total Hours	Cleaned Hours
Cicipu	5.66	3.09
Mocho'	7.26	4.21
Toratán	22.84	11.15
Ulwa	3.25	2.83
Upper Napo Kichwa	13.19	6.97

Table 6: Total archived and cleaned hours of audio for all languages used in the study.

C Error Rates

We consider four phonological categories:

$$C \in \{\text{Tone, Nasality, V\_length, C\_length}\}.$$

Over the entire dataset, we record:

- S_C: the total *substitution* errors for category C,
- D_C: the total *deletion* errors for category C,
- I_C: the total *insertion* errors for category C,
- L_C: the total *reference tokens* exhibiting category C (e.g. `tone_labels` for tone, `total_vowels` for vowel length, etc.).

We then define the total errors and error rate for each category C as follows:

$$E_C = S_C + D_C + I_C \quad \text{and} \quad \text{ErrorRate}_C = \frac{E_C}{L_C}.$$

For example, if C = tone then L_C = tone_labels (the number of reference tokens with at least one tone diacritic). Similarly, if C = vowel_length then L_C = total_vowels (the total vowel tokens in the reference).